Snowballing Effects in Preferential Attachment: The Impact of The Initial Links

Huanyang Zheng and Jie Wu
Department of Computer and Information Sciences, Temple University, USA
Email: {huanyang.zheng, jiewu}@temple.edu

***Abstract*—This paper studies the node degree snowballing effects (i.e., degree growth effects) in the age-sensitive preferential attachment model, where nodes are iteratively added one by one to a growing network. Upon entering the network, each new node connects to a suitably chosen set of existing nodes, while the attachment probability for an existing node to get connected depends on both its node degree and age difference. We are interested in accelerating the node degree snowballing effects through the impact of the initial links. If a new node enters the growing network with more initial links (a larger degree), it could attract many more links from the later nodes, and thus, its degree snowballs faster. We find that the initial links are only impactful when neither the node degree nor the age difference dominates the attachment probability. In that case, the relationship between the ratio of the additional initial link and the gain ratio of the eventual node degree is shown to include two stages (linear stage and diminishing return stage). Applications of our work involve citation networks and online social networks. For example, in citation networks, we answer the question that whether an author can attract additional citations through self-citations. Finally, real data-driven experiments verify the accuracies of our results, which cast some new light in real-world growing networks.**

Keywords*—Node degree snowballing effects, preferential attachment model, percolation phenomena.

I. Introduction

One of the most impressive recent discoveries in the field of network evolution is the observation that a number of large growing networks are scale-free [1–3]. Their key feature is that the node degree distributions have a power-law form [4, 5]. Typical scale-free networks include the citation networks, the online social networks, the World Wide Web, and so on. The preferential attachment model is one of the most acknowledged models for explaining the formation of scale-free networks [6]. In this model, nodes are iteratively added one by one to a growing network (one new node per time unit). Upon entering the network, each new node connects to a suitably chosen set of existing nodes, while the attachment probability for an existing node to get connected is proportional to its degree. Therefore, the existing node with a large degree is preferentially attached, resulting in the *degree snowballing effect* (i.e., degree growth effect), in which the rich get richer.

In this paper, we are interested in accelerating such degree snowballing effects through the impact of the initial links. The initial links of a node are the links set by that node at the time when it enters the growing network. If a new node enters the growing network with more initial links, it could attract many more links from the later nodes, and thus its degree snowballs faster. While the impact of the initial links remains unexplored, it has important applications as follows.

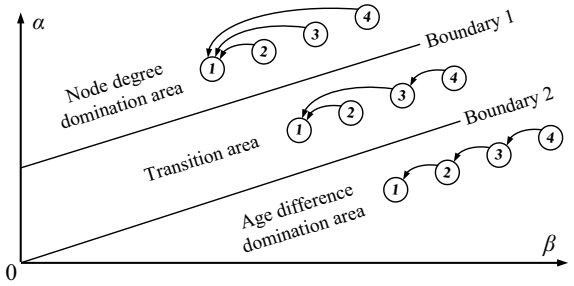


Fig. 1. The age-sensitive preferential attachment.

(1) In citation networks, we are more likely to cite papers with high citations than that with low citations. Then, if an author cites his/her own papers (self-citation), is it possible for the papers of this author to gain extra citations at a later time by the snowballing effect? (2) In online social networks such as Facebook and Twitter, business pages want to attract more followers, as to propagate the product information for sales. Then, if a business page makes an advertisement (e.g., Facebook page promotion [7]), how many additional followers can this business page attract at a later time by the snowballing effect? The impact of the initial links is explored in our study, which is critical for the development of the network science.

To be more realistic, here we study the degree snowballing effects in *age-sensitive preferential attachment models*, where the attachment probability is age-sensitive [8]. For example, in citation networks, we prefer to cite recent papers more than old papers. Specifically, we consider that the attachment probability for an existing node to get connected is proportional to $d^{\alpha} \cdot \Delta t^{-\beta}$. Here, d is the degree of the existing node, while Δt is the age difference (also entry time difference) between the new node and the existing node. α and β ($\alpha > 0$ and $\beta > 0$) are parameters obtained by existing estimators [9]. The age-sensitive model can reduce to the classic model when $\alpha = 1$ and $\beta = 0$. Then, in terms of the attachment probability, there exists a tradeoff between the attractiveness brought by the node degree and the repulsiveness brought by the age difference. Although older nodes have larger degrees, they may not attract more links from the new nodes, due to the larger age difference. An example is shown in Fig. 1, where the nodes enter the growing network one by one (following their IDs). Upon entering the network, the node connects to existing nodes according to the node degree and the age difference. It can be seen that the resulting network structure of the age-sensitive preferential attachment model depends on α and β (age difference domination area, transition area, and node degree domination area).

The snowballing effects in age-sensitive preferential attachment models are more intriguing and challenging. With respect to α and β, how does the impact of the initial links vary? Since those three areas in Fig. 1 result in different network structures, the impact of the initial links should be qualitatively different. Moreover, is the amount of the initial links important? While a small amount of the initial links leads to a limited change, a large amount of the initial links may lead to a big change.

Our results and contributions are summarized as follows:

- Percolation phenomena are found in the age-sensitive preferential attachment models. Boundaries 1 and 2 in Fig. 1 are $\alpha = \beta + 1.5$ and $\alpha = \beta$, respectively. We show that the initial links are not impactful in the node degree domination area and the age difference domination area.
- We show that the initial links are only impactful in the transition area of $\beta \leq \alpha \leq \beta + 1.5$. In that case, the impact of the initial links is found to have two stages (linear stage and diminishing return stage). We further show that the initial links are most impactful, when the corresponding growing network lies in the "middle" of the transition area.
- Accuracies of our theoretical results are verified. The degree snowballing effects are observed in the real-world citation network and online social network.

The remainder of this paper is organized as follows. Section II is the related work. In Section III, we set up the model and formulate the problem. In Section IV, the impact of the initial links is analyzed. Section V includes the experiments. Section VI shows the conclusion. Proofs are presented in the Appendix.

II. Related Work

The classic preferential attachment model was proposed in [10] with a well-studied body of knowledge in the network science. The aging effects have been observed. For example, Wang et al. [11] studied the predictability of the citation patterns with respect to different time slots. Zhao et al. [12] explored the multi-scale dynamics of time-sensitive information propagations. Authors in [8, 13] studied the scale-free properties in the age-sensitive preferential attachment models, in terms of the degree distributions and the clustering properties. The aging effects are preliminarily explored in the citation networks [14], and then are found in the online social networks [15], the World Wide Web [16], the recommendation systems [17], and so on. These works mainly focus on the scale-free properties, where the node degree distribution follows power-law form. In contrast, we explore the degree snowballing effects in growing networks, which are completely novel.

The other existing findings that are highly related to the snowballing effects include the rich-get-richer phenomenon, the "Matthew effect" [18], and the cumulative advantage [19]. Although they have been empirically confirmed for a long time with respect to the economic market, quantitative studies have not been conducted for the citation networks and the online social networks. For example, Kumar et al. [20] studied the equilibrium states of two-sided market evolution through an empirical analysis on the cumulative capital advantage. Braha et al. [21] simulated the corporate competition in the preferential attachment model with respect to the snowball effect. Kas et al. [22] studied the structures and statistics of citation networks. However, they did not consider the impact of the initial links, which is explored in this paper.

III. Model and Problem Formulation

A. Preferential Attachment Model

In the preferential attachment model [6], nodes are iteratively added one by one to a growing network (one new node per time unit). The node added at the time s is denoted as N_s, while the current time is denoted as t ($t \geq s$). The age of a node is its existing time in the growing network, i.e., the age of the node N_s is $t - s$. Upon entering the network, each new node connects to a suitably chosen set of existing nodes, while the attachment probability for an existing node to get connected is proportional to $d^{\alpha} \cdot \Delta t^{-\beta}$. Here, d is the degree of the existing node, while Δt is the age difference (also entry time difference) between the new node and the existing node. α and β ($\alpha > 0$ and $\beta > 0$) are parameters obtained by existing estimators [9]. The initial links of a node are the links set by that node at the time when it enters the growing network. We assume that each new node sets m new links to the existing nodes. The links are directional, while the node degree is the summation of its in-degree and out-degree. Let $d(s, t)$ denote the expected degree of the node s at the time t ($t \geq s$), while the initialization condition is $d(s, s) = m$.

Since an existing node will get attached by later nodes, the degree of an existing node snowballs with respect to the time. However, in terms of the snowballing speed, there exists a tradeoff between the attractiveness brought by the node degree (with a larger α being more attractive) and the repulsiveness brought by the age difference (with a larger β being more repulsive). Although older nodes have larger degrees, they may not attract more links from the new nodes, due to the larger age differences. As previously shown in Fig. 1, the resulting network structure of the age-sensitive preferential attachment model depends on α and β (age difference domination area, transition area, and node degree domination area).

B. Problem Formulation

In this paper, we study the impact of the initial links in the age-sensitive preferential attachment models. While a normal node enters the growing network with only m links, we focus on a particular node that enters the network with additional m' links ($m + m'$ links in total), as to observe the impact of the initial links. Since a larger degree means a larger attachment probability, the additional initial links can accelerate the degree snowballing effects for the nodes in the growing network. Our study has important applications as follows.

- In citation networks, we are more likely to cite papers with a high number citations than that with a low number of citations. Then, if an author cites his/her own papers (self-citation), is it possible for the papers of this author to gain extra citations at a later time? In this scenario, m and m' represent the average paper citations and the number of self-citations, respectively.
- In online social networks such as Facebook and Twitter, business pages want to attract more followers, as to

propagate the product information for sales. Then, if a business page makes an advertisement (e.g., Facebook page promotion [7]), how many additional followers can this business page attract at a later time by the snowballing effect? Here, m' can be interpreted as the number of followers attracted by the advertisement.

For the simplicity of the following analysis, we define the *initial rate* (r_i) as the ratio of the additional initial links to the normal initial links (i.e., $r_i = m'/m$). A larger initial rate means that the corresponding node has a larger initial degree.

If the node N_s enters the network with an additional m' links, then we use $d'(s,t)$ to denote its expected degree at the time t ($t \geq s$). Its initialization condition is $d'(s,s) = m+m'$. We are interested in the ratio of the node degree gain brought by the additional initial links, which is defined as the *gain rate* (denoted by r_g). In other words, we have:

$$r_g = \frac{d'(s,t) - d(s,t)}{d(s,t)} \tag{1}$$

The objective of this paper is to study *the relationship between the initial rate and the gain rate*, which represents the impact of the initial links in the growing networks. A larger initial rate should bring a non-smaller gain rate. We also want to study how this relationship changes with respect to the parameters α, β, s, and t. Note that, α and β indicate the attractiveness brought by the node degree and the repulsiveness brought by the age difference, respectively. Therefore, the values of α and β are also important for the initial links to be impactful in the corresponding growing network. Meanwhile, s indicates the time for introducing the additional initial links. Our analyses are shown in the next section.

IV. Snowballing Effects in Age-Sensitive Preferential Attachment

In this section, we study the relationship between the initial rate and the gain rate, as to understand the impact of the initial links. First, we review the classic preferential attachment model. Then, we look into the snowballing effects within the node degree domination area and the age difference domination area of the age-sensitive model, respectively. Finally, we show the snowballing effects within the transition area.

A. Classic Preferential Attachment

In the classic preferential attachment model [6], the attachment probability for an existing node to get connected is only proportional to its degree ($\alpha = 1$ and $\beta = 0$). Let us start with the case for a normal node that enters the network with m links. Then, when a new node enters the network at the time $t+1$, the attachment probability for the node N_s to get connected is:

$$\frac{d(s,t)}{\sum_{s=1}^{t} d(s,t)} = \frac{d(s,t)}{2mt} \tag{2}$$

The denominator $\sum_{s=1}^{t} d(s,t)$ is the total degree, which is the normalization factor in Eq. 2. The total degree is $2mt$, since there are t nodes in the network and each node has brought m links. We assume that the attachment processes for the m links are independent of each other, and thus the expected degree gain of the node N_s is $m \times \frac{d(s,t)}{2mt} = \frac{d(s,t)}{2t}$. In other words, we have the following equation:

$$d(s,t+1) = d(s,t) + m \times \frac{d(s,t)}{2mt} = \frac{2t+1}{2t} d(s,t) \tag{3}$$

If we do the recursion in Eq. 3, then we can get:

$$\begin{aligned} d(s,t) &= \frac{2t-1}{2t-2} \times \frac{2t-3}{2t-4} \times \cdots \times \frac{2s+1}{2s} \times d(s,s) \\ &= \exp\Big\{ \ln \frac{2t-1}{2t-2} + \cdots + \ln \frac{2s+1}{2s} \Big\} \times d(s,s) \\ &\approx \exp\Big\{ \frac{1}{2t-2} + \cdots + \frac{1}{2s} \Big\} \times d(s,s) \\ &\approx \exp\Big\{ \frac{1}{2} \ln \frac{t}{s} \Big\} \times d(s,s) = m\sqrt{\frac{t}{s}} \end{aligned} \tag{4}$$

In Eq. 4, we have used the approximations of $\ln \frac{2s+1}{2s} \approx \frac{1}{2s}$ and $\sum_{x=s}^{t-1} \frac{1}{2x} \approx \int_s^t \frac{1}{2x} \mathrm{d}x$. Eq. 4 implies that the node degree has a square-root growth with respect to the ratio of the current time to the node entry time. Similar to Eq. 4, if the node N_s enters the network with an additional m' links, we can get:

$$d'(s,t) \approx \exp\Big\{ \frac{1}{2} \ln \frac{t}{s} \Big\} \times d'(s,s) = (m+m')\sqrt{\frac{t}{s}} \tag{5}$$

Eqs. 4 and 5 mean that *the gain rate equals the initial rate* (i.e., $r_g = r_i$) in the classic preferential attachment model. Here we have assumed that the number of additional initial links is small (i.e., $m' \ll 2mt$). However, the relationship of $r_g = r_i$ is uncommon in real-world growing networks, since the prerequisite that the attachment probability is only proportional to the degree may not be true.

In the following three subsections, we will discuss the snowballing effects in the age-sensitive preferential attachment model, where the attachment probability is determined by both the node degree and the age difference. As previously mentioned, the attachment probability is proportional to $d^{\alpha} \cdot \Delta t^{-\beta}$. The tradeoff between the attractiveness brought by the node degree and the repulsiveness brought by the age difference divides the resulting network structure into three areas (age difference domination area, node degree domination area, and transition area). Each of the three following subsections corresponds to one of those three areas.

B. Age Difference Domination Area

In this subsection, we study the snowballing effects in the age-sensitive preferential attachment model, in which the age difference dominates the attachment probability. In other words, the attractiveness brought by the node degree is much smaller than the repulsiveness brought by the age difference. To study the snowballing effects, we first need to clarify the boundary of this area, as shown in the following theorem:

***Theorem** 1:* When $\alpha < \beta$, the first node will attract a finite number of links, with respect to the network growth.

The proof of Theorem 1 is shown in Appendix A. The basic idea of the proof is to show that, when $\alpha < \beta$, the first node is much less attractive than a younger node for a new node to attach. The insight behind Theorem 1 is that the age difference dominates the attachment probability, where the new nodes are more intended to link to the younger nodes. At this time, even

if an old node has a very high degree, it will not be further attached to by the new nodes. The resulting network structure for this case is illustrated in Fig. 2(a), where the nodes connect to each other one by one following their entry times. As for the snowballing effects, we have:

***Theorem** 2:* When $\alpha < \beta$, for the node N_s that enters the growing network at the time s, it needs at least $\Omega((t-s)^{\beta/\alpha})$ additional initial links to keep its attractiveness for nodes that enter the growing network at the time t.

The proof of Theorem 2 is shown in Appendix B. The basic idea of the proof is to show that, when $\alpha < \beta$, the node N_s needs many additional initial links to resist the dominated repulsiveness brought by the age difference. The insight behind Theorem 2 is that the initial links in the growing network with $\alpha < \beta$ are not impactful, since the initial links are wasted on resisting the dominated aging effects. In other words, the gain rate is close to zero, unless we have a very large initial rate (basically impossible for real-world growing networks).

C. Node Degree Domination Area

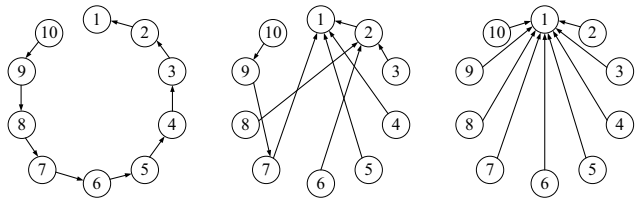
In this subsection, we study the snowballing effects in the age-sensitive preferential attachment model, in which the node degree dominates the attachment probability. In other words, the attractiveness brought by the node degree is much larger than the repulsiveness brought by the age difference. Similarly, we first need to clarify the boundary of this area, as shown in the following theorem:

***Theorem** 3:* When $\alpha > \beta + 1.5$, the first node will attract an infinite number of links, with respect to the network growth. The first node N_1 has a degree of $\Theta(t)$.

The proof of Theorem 3 is shown in Appendix C. The basic idea of the proof is to show that, when $\alpha > \beta + 1.5$, the first node is much more attractive than the remaining nodes for a new node to attach. The insight behind Theorem 3 is that the degree dominates the attachment probability, where the new nodes are more likely to attach to the oldest node. At this time, younger nodes will not be further attached by the new nodes, while the first node has a degree of $\Theta(t)$. In other words, the first node monopolizes the majority of the links. The resulting network structure for this case is illustrated in Fig. 2(c). Note that, the classic model with $\alpha = 1$ and $\beta = 0$ lies under the boundary of $\alpha = \beta + 1.5$, which has a qualitative difference with models in the node degree domination area. The first node only has a degree of $\Theta(\sqrt{t})$ when $\alpha = 1$ and $\beta = 0$. As for the snowballing effects, we have:

***Theorem** 4:* When $\alpha > \beta + 1.5$, for the node N_s that enters the growing network at the time s, it needs at least $\Omega(s^{\alpha-\beta})$ additional initial links to keep its attractiveness for later nodes.

The proof of Theorem 4 is shown in Appendix D. The basic idea of the proof is to show that, when $\alpha > \beta + 1.5$, the node N_s needs many additional initial links to compete with the node N_1, in terms of attracting the new attachments. Theorem 3 states that the first node N_1 has a degree of $\Theta(s)$ at the time s. The insight behind Theorem 4 is that a large number of additional initial links is needed to break the link monopoly of the node N_1. In other words, the gain rate is also close to zero, unless a very large initial rate is used. Therefore, the initial links in the growing network with $\alpha > \beta + 1.5$ are not impactful on the eventual node degree, the result of which is similar to that for the age difference domination area.



(a) $\alpha < \beta$ (b) $\beta \le \alpha \le \beta + 1.5$ (c) $\alpha > \beta + 1.5$

Fig. 2. The percolation phenomena in the age-sensitive preferential attachment ($m = 1$ and $t = 10$). The ID of a node is its entry time.

D. Transition Area

In the previous two subsections, Theorems 1 and 3 show that $\alpha = \beta$ and $\alpha = \beta + 1.5$ are two boundaries for the percolation phenomena in the age-sensitive preferential attachment models. When $\alpha < \beta$ or $\alpha > \beta + 1.5$, the resulting network structure turns out to be simplex. The resulting network structure for the transition area of $\beta \le \alpha \le \beta + 1.5$ is more complex. An example for the transition area is illustrated in Fig. 2(b). Meanwhile, Theorems 2 and 4 show that the initial links are not impactful in the age difference domination area and the node degree domination area, since the initial links are wasted to resist the dominated power.

In this subsection, we discuss the snowballing effects in the transition area of $\beta \le \alpha \le \beta + 1.5$, based on [8]. Let us start with the case for a normal node that enters the network with m links. Similar to Eq. 3, we have:

$$d(s,t+1) = d(s,t) + m \times \frac{d(s,t)^{\alpha}(t-s)^{-\beta}}{\sum_{s=1}^{t} d(s,t)^{\alpha}(t-s)^{-\beta}} \quad (6)$$

When $\alpha = 1$ and $\beta = 0$, Eq. 6 is reduced to Eq. 3 (the classic model). Eq. 6 can also be written in the continuous form:

$$\frac{\partial d(s,t)}{\partial t} = m \times \frac{d(s,t)^{\alpha}(t-s)^{-\beta}}{\int_1^t d(s,t)^{\alpha}(t-s)^{-\beta}\mathrm{d}s} \quad (7)$$

Since Eq. 7 is very complex, we consider the node degree to be scaling ($d(s,t) \equiv d(s/t)$) [8]. In other words, the node degree is considered as a function of s/t . For notation simplicity, we set $\xi = s/t$. Then, Eq. 7 can be rewritten as:

$$\frac{1}{d(\xi)^{\alpha}} \times \frac{\mathrm{d}d(\xi)}{\mathrm{d}\xi} = \frac{-1}{\xi(1-\xi)^{\beta}} \frac{m}{\int_0^1 d(\xi)^{\alpha}(1-\xi)^{-\beta}\mathrm{d}\xi} \quad (8)$$

If we do the integral in Eq. 8, we can get:

$$\frac{d(\xi)^{1-\alpha} - d(1)^{1-\alpha}}{1-\alpha} = \frac{m\int_1^{\xi} \frac{-1}{\xi(1-\xi)^{\beta}}\mathrm{d}\xi}{\int_0^1 d(\xi)^{\alpha}(1-\xi)^{-\beta}\mathrm{d}\xi} \quad (9)$$

When $\alpha \to 1$ and $\beta = 0$, Eq. 9 reduces to $\ln\frac{d(\xi)}{d(1)} = \frac{-1}{2}\ln\xi$ that is consistent with Eq. 4, i.e., $d(s,t) = m\sqrt{t/s}$. This is because $d(\xi)^{1-\alpha} \approx e^{(1-\alpha)\ln d(\xi)} \approx 1 + (1-\alpha)\ln d(\xi)$, when $\alpha \to 1$. The result in Eq. 9 can be rewritten as:

$$\begin{aligned} d(\xi) &= \left[m^{1-\alpha} + \frac{(1-\alpha)m\int_1^{\xi} \frac{-1}{\xi(1-\xi)^{\beta}}\mathrm{d}\xi}{\int_0^1 d(\xi)^{\alpha}(1-\xi)^{-\beta}\mathrm{d}\xi}\right]^{\frac{1}{1-\alpha}} \\ &\approx m \times [1 + (1-\alpha)C(\alpha,\beta,\xi)]^{\frac{1}{1-\alpha}} \end{aligned} \quad (10)$$

Eq. 10 is approximated by using $d(\xi) = m\xi^{-1/2}$ to calculate the normalization factor. This approximation is feasible, since it can represent the degree distribution in the transition area. Meanwhile, the $C(\alpha, \beta, \xi)$ in Eq. 10 is:

$$C(\alpha, \beta, \xi) = \frac{\int_1^{\xi} \frac{-1}{\xi(1-\xi)^{\beta}} \mathrm{d}\xi}{\int_0^1 \xi^{-\alpha/2}(1-\xi)^{-\beta} \mathrm{d}\xi} \tag{11}$$

Similar to Eq. 10, if the node N_s enters the network with m' additional links, then we can get:

$$\begin{aligned} d'(\xi) &= \left[(m+m')^{1-\alpha} + \frac{m\int_1^{\xi} \frac{-1}{\xi(1-\xi)^{\beta}} \mathrm{d}\xi}{\int_0^1 d(\xi)^{\alpha}(1-\xi)^{-\beta} \mathrm{d}\xi}\right]^{\frac{1}{1-\alpha}} \\ &\approx m \times [(1+\frac{m'}{m})^{1-\alpha} + (1-\alpha)C(\alpha, \beta, \xi)]^{\frac{1}{1-\alpha}} \end{aligned} \tag{12}$$

In Eq. 12, we have assumed that m' is small enough with respect to the normalization factor of $\int_0^1 d(\xi)^{\alpha}(1-\xi)^{-\beta} \mathrm{d}\xi$. Combining Eqs. 10, 11, and 12, the relationship between the initial rate and the gain rate can be obtained:

$$r_g = \left[\frac{(1+r_i)^{1-\alpha} + (1-\alpha)C(\alpha, \beta, \xi)}{1 + (1-\alpha)C(\alpha, \beta, \xi)}\right]^{\frac{1}{1-\alpha}} - 1 \tag{13}$$

Note that, $C(\alpha, \beta, \xi)$ can be regarded as a constant with respect to r_i. Meanwhile, we have $C(1, 0, \xi) = \frac{1}{2}\ln\xi$. Further analysis on Eq. 13 shows the following theorem:

***Theorem** 5:* When $|(1-\alpha)C(\alpha, \beta, \xi)| \gg 1$, the gain rate is close to zero (i.e., $r_g \approx 0$). When $|(1-\alpha)C(\alpha, \beta, \xi)| \ll 1$, the relationship between the initial rate and the gain rate satisfies $r_g \approx (1+r_i)e^{-C(\alpha,\beta,\xi)} - 1$.

The proof of Theorem 5 is shown in Appendix E. Theorem 5 shows three intriguing properties for the impact of the initial links, while the first property is that there is a prerequisite for the initial links to be impactful. This threshold results from the fact that either the dominated attractiveness brought by the node degree or the dominated repulsiveness brought by the age difference can weaken the impact of the initial links. This is similar to the case in the node degree domination area or the age difference domination area. If this threshold is satisfied, then the gain rate increases linearly with respect to the initial rate. However, note that Theorem 5 is derived under the assumption that m' is small enough with respect to the normalization factor in Eq. 7. If the initial rate is very large ($r_i \in \Omega(\int_0^1 \xi^{-\alpha/2}(1-\xi)^{-\beta} \mathrm{d}\xi)$), the gain rate has a diminishing return effect with respect to the initial rate. This is because the total links in the network are limited: A node cannot attract more than $2mt$ links, no matter how many additional initial links are given. Therefore, the relationship between r_g and r_i has two stages as shown in Fig. 3 (denoted as the linear stage and the diminishing return stage).

The second property revealed by Eq. 13 is that the initial links are most impactful when the attractiveness brought by the node degree and the repulsiveness brought by the age difference cancel each other out. This is because the threshold of $|(1-\alpha)C(\alpha, \beta, \xi)|$ will be small in such a case. The first node will not monopolize the majority of the links, while the aging effect will not prevent the new entering node from connecting to old nodes. The links from new entering nodes in the growing network will evenly connect to both the old and the young nodes. Actually, the classic model with $\alpha = 1$ and $\beta = 0$ is such a case (as previously shown in Eq. 5), where the gain rate is strictly linear with respect to the initial rate. This is because the threshold of $|(1-\alpha)C(\alpha, \beta, \xi)|$ becomes zero for $\alpha = 1$ and $\beta = 0$. Moreover, this threshold can be used to estimate whether the initial links are impactful in the corresponding growing network or not. The initial links are most impactful, when the corresponding growing network lies in the "middle" of the transition area.

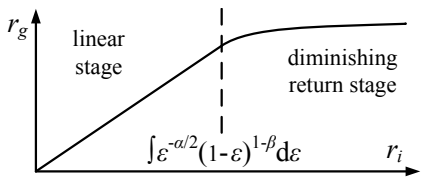


Fig. 3. The two-stage relationship between the gain rate (r_g) and the initial rate (r_i), when $|(1-\alpha)C(\alpha, \beta, \xi)| \ll 1$.

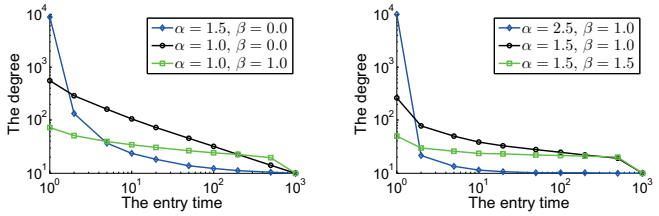
The third property revealed by Eq. 13 is on the impact of the time period. Note that $-C(\alpha, \beta, \xi)$ decreases monotonously with respect to ξ ($\xi = s/t$). Meanwhile, the slope of the linear stage is approximately $e^{-C(\alpha,\beta,\xi)}$ (a smaller ξ brings a larger slope). The insight is that the initial links become more impactful with respect to a longer time period (the snowballing effect becomes more significant). Further explorations on the relationship between $C(\alpha, \beta, \xi)$ and ξ (representing the impact of the time) will be our future work.

V. Experiments

In this section, we first set up the age-sensitive preferential attachment model to verify the accuracies of our theoretical results. Then, the snowballing effects are studied in several real-network datasets. The experimental results are shown from different perspectives to provide insightful conclusions.

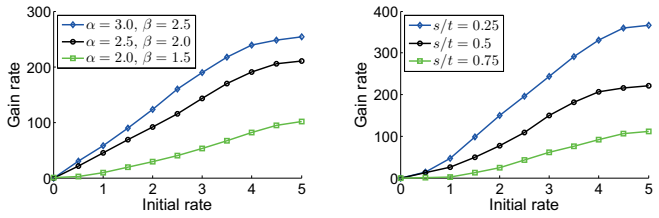
A. Accuracy Verifications on Theoretical Results

In this subsection, we verify the accuracy of our theoretical results for the age-sensitive preferential attachment model, which has a duration of 1,000 time slots (i.e., t =1,000). Upon each time slot, one new node will enter the network with 10 new connections to the existing nodes (i.e., $m = 10$). First, we check the expected node degree distributions with respect to the node entry time s. The results, which are averaged over 1,000 times, are shown in Fig. 4 as a log-log plot. Figs. 4(a) and 4(b) show two scenarios with different values of α and β. It can be seen that, when $\alpha = \beta + 1.5$, the first node attracts almost all the new links from the later nodes (there are $m \times t$ =10,000 links in total). At this time, the expected node degree decays quickly with respect to the node entry time, due to the overpowered attractiveness brought by the node degree. On the other hand, when $\alpha = \beta$, all the nodes tend to have a degree of $O(m)$, since the overpowered repulsiveness brought by the age difference only enables new nodes to connect to the most recent nodes. It can also be seen that, when neither the node degree nor the age difference dominates the attachment probability ($\beta \leq \alpha \leq \beta + 1.5$), the resulting network structure is more complex, due to the fact that the new nodes will connect to both old and young existing nodes. The experiment



(a) Scenario 1. (b) Scenario 2.

Fig. 4. The node degree distributions with respect to the node entry time.



(a) $s/t = 0.5$. (b) $\alpha = 2.5$, $\beta = 2.0$.

Fig. 5. The node degree snowballing effects.

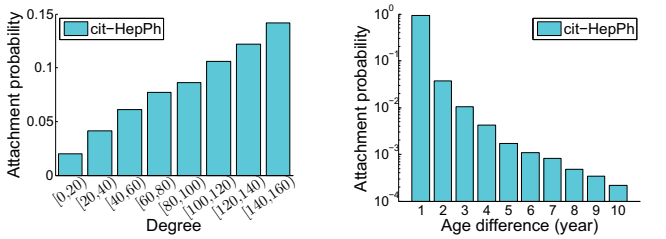
verifies the existence of the percolation phenomena in the age-sensitive preferential attachment model.

The snowballing effects in the transition area are also experimentally studied. Fig. 5(a) shows the relationship between the initial rate and the gain rate for the node that enters the network at the 500^{th} time slot (i.e., $s/t = 0.5$), under three different settings of α and β. Each of the curves in Fig. 5(a) clearly has two stages (linear stage and diminishing return stage) as previously analyzed. Then, Fig. 5(b) shows the impact of the time, under $\alpha = 2.5$ and $\beta = 2.0$. It can be seen that a smaller s/t leads to a larger gain rate, since more nodes will enter the network after the node N_s. These experimental results confirm that our theoretical results in Eq. 13 are accurate. In the following two subsections, we will further verify the node degree snowballing effects in real data-driven experiments (the citation network and the online social network).

B. Citation Network

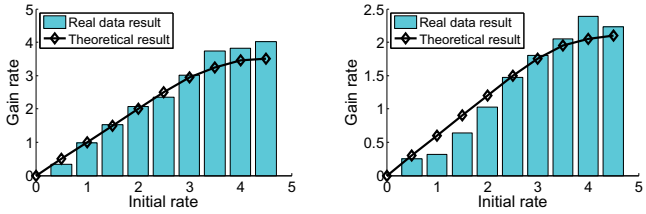
In this subsection, we conduct real data-driven experiments to verify the snowballing effects in citation networks. Citation networks are classic growing networks, where papers serve as nodes in the network. New papers enter the network as time goes by. If a paper i cites a paper j, then the network contains a directed link from i to j. It is common sense that authors generally prefer to cite papers with a high number of citations (compared to papers with a low number of citations), as well as recent papers (compared to old papers). Currently, it is well-known that the classic preferential attachment model can decently explain the formation of citation networks [6].

Our experiments use the real dataset [23] of the Arxiv high energy physics phenomenology citation network (denoted as cit-HepPh). This dataset covers the papers published over the period from January 1993 to April 2003. 34,546 papers (nodes) and 421,578 citations (links) are involved in this dataset. On average, a new paper enters the growing network every 0.12 days (i.e., the time unit for adding one new node to the growing network). The relationships between the attachment probability



(a) The relationship between attachment probability and node degree. (b) The relationship between attachment probability and age difference.

Fig. 6. The age-sensitive preferential attachment in cit-HepPh.



(a) Papers published in 1995. (b) Papers published in 1998.

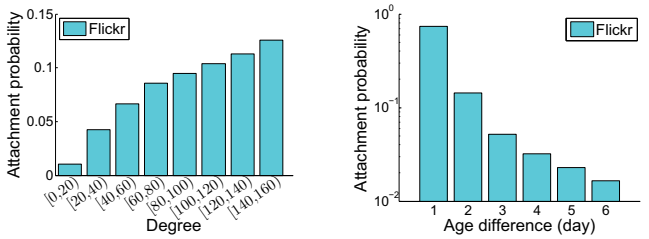
Fig. 7. The node degree snowballing effects in cit-HepPh.

and the node degree (or age difference) are shown in Fig. 6. It verifies the feasibility of the assumption that the attachment probability is proportional to $d^{\alpha} \cdot \Delta t^{-\beta}$. Then, we use the maximum likelihood to estimate the exponents α and β in the attachment probability. On average, we get $\alpha = 0.91$ and $\beta = 1.2 \times 10^{-3}$ as the exponents in the attachment probability. Note that β is small, due to the ground truth that we sometimes cite a paper from 10 years ago (more than 30,000 time units).

To study the snowballing effects, papers published in the years 1995 and 1998 are analyzed. If a paper has more than an average number of citations in its first publication year, the additional portion is regarded as its initial rate. The final number of citations of that paper (in the year 2003) and the average number of citations are used to calculate the gain rate of that paper. The result is shown in Fig. 7, where we have $\alpha = 0.91$ and $\beta = 1.2 \times 10^{-3}$. It can be seen that the relationship between the initial rate and the gain rate has two stages in Fig. 7 (i.e., real data result). It is consistent with Eq. 13, which is denoted as the theoretical result in Fig. 7. It can also be seen that, the initial links are more impactful for earlier papers. For the same initial rates, papers published in 1995 have higher gain rates than did those in 1998.

C. Online Social Network

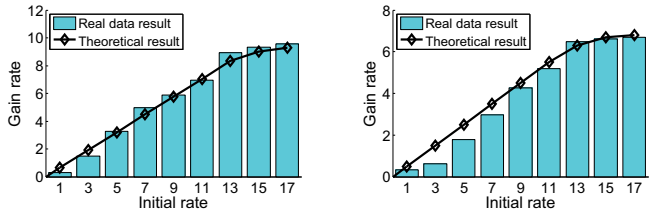
In this subsection, we conduct real data-driven experiments to verify the snowballing effects in the online social networks, which are platforms for users to build social relations among other users. Users in the network share news, stories, and photos with each other. Social network sites are web-based services that allow individuals to create a public profile, to create a list of users with whom to share connections, and view and cross the connections within the system. Online social networks are growing networks, where users (i.e. nodes) enter the network one by one. If a new user i follows an existing user j, then the network contains a directed link from i to j. Users are more likely to follow popular users with high degrees, as



(a) The relationship between attachment probability and node degree.

(b) The relationship between attachment probability and age difference.

Fig. 8. The age-sensitive preferential attachment in Flickr.



(a) Users entered on Monday.

(b) Users entered on Wednesday.

Fig. 9. The node degree snowballing effects in Flickr.

well as contemporary users with smaller age differences. It is well-known that the classic preferential attachment model can also be applied to the online social networks [6].

In the experiments, we use the Flickr dataset [24]. Flickr is an online social network for sharing photos. Key features of Flickr not initially present are tags, marking photos as favorites, group photo pools, and interestingness, for which a patent is pending. This dataset covers all new users in the period from November 2006 to May 2007, including 167,527 users (nodes) and 526,874 follower-followee relationships (links). On average, a new user enters the growing network every 0.04 days (i.e., the time unit for adding one new node to the growing network). The relationships between the attachment probability and the node degree (or age difference) are shown in Fig. 8. It again verifies the feasibility of the assumption that the attachment probability is proportional to $d^{\alpha} \cdot \Delta t^{-\beta}$. We also use the maximum likelihood to estimate α and β. On average, it turns out that we have $\alpha = 0.89$ and $\beta = 1.3 \times 10^{-4}$ in this dataset. Note that β is also small in this dataset, due to the ground truth that new users sometimes follow old users (one month is about 750 time units).

To study the snowballing effects, we focus on the users who entered the network in the first week of April. These users are selected, since they have complete records in the dataset (while the records of some other users may be missing). If a user has above average connections in his/her first day, the additional portion is regarded as its initial rate. Meanwhile, the final number of connections of that user (at the end of this week) and the average number of connections are used to calculate the gain rate of that user. The result is shown in Fig. 9, where we have $\alpha = 0.89$ and $\beta = 1.3 \times 10^{-4}$. It can be seen that the relationship between the initial rate and the gain rate should also have two stages, as shown in Fig. 9 (i.e., real data result). Although our theoretical result in Eq. 13 has a little overestimation in Fig. 9(b), it is basically accurate for this dataset. The initial links are also more impactful for earlier users in this dataset. For the same initial rates, users who entered on Monday have higher gain rates than do those who entered on Wednesday.

VI. Conclusion

In this paper, we study the node degree snowballing effects in the age-sensitive preferential attachment model, where the attachment probability depends on both the node degree and the age difference. We are interested in accelerating such degree snowballing effects through the impact of the initial links. Our study answers the question 'how many additional citations can an author obtain through self-citations?' The percolation phenomena are found in the age-sensitive preferential attachment model: the initial links are only impactful in the transition area, where neither the node degree nor the age difference dominates the attachment probability. In that case, we show that the relationship between the initial rate and the gain rate has two stages (linear stage and diminishing return stage). Real data-driven experiments in the citation network and the online social network verify the accuracies of our theoretical results, which cast some new light on the impact of the initial links in real-world growing networks. Further explorations on the impact of the time will be our future work.

VII. Acknowledgments

This work is supported in part by NSF grants CNS 149860, CNS 1461932, CNS 1460971, CNS 1439672, CNS 1301774, ECCS 1231461, ECCS 1128209, and CNS 1138963.

References

[1] H. Jeong, Z. Neda, and A.-L. Barabási, "Measuring preferential attachment in evolving networks," *Europhysics Letters*, vol. 61, no. 4, pp. 567–572, 2007.

[2] E. Bulut and B. K. Szymanski, "Constructing limited scale-free topologiesover peer-to-peer networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 4, pp. 919–928, 2014.

[3] A. Mahanti, N. Carlsson, M. Arlitt, and C. Williamson, "A tale of the tails: Power-laws in internet measurements," *IEEE Network*, vol. 27, no. 1, pp. 59–64, 2013.

[4] J. Tan, B. Swapna, and N. B. Shroff, "Retransmission delays with bounded packets: Power-law body and exponential tail," *IEEE/ACM Transactions on Networking*, vol. 22, no. 1, pp. 27–38, 2014.

[5] M. Ripeanu, I. Foster, and A. Iamnitchi, "Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design," *IEEE Internet Computing*, 2002.

[6] M. E. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, p. 25102, 2001.

[7] https://www.facebook.com/advertising.

[8] S. N. Dorogovtsev and J. F. Mendes, "Scaling properties of scale-free evolving networks: Continuous approach," *Physical Review E*, vol. 63, no. 5, p. 56125, 2001.

[9] http://tuvalu.santafe.edu/%7Eaaronc/powerlaws/.

[10] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[11] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, pp. 127–132, 2013.

[12] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao, "Multi-scale dynamics in a massive online social network," in *Proceedings of ACM IMC 2012*, pp. 171–184.

[13] Y. Wu, T. Z. Fu, and D. M. Chiu, "Generalized preferential attachment considering aging," *Journal of Informetrics*, vol. 8, no. 3, pp. 650–658, 2014.
[14] M. Wang, G. Yu, and D. Yu, "Effect of the age of papers on the preferential attachment in citation networks," *Physica A*, vol. 388, no. 19, pp. 4273–4276, 2009.
[15] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of ACM IMC 2007*, pp. 1–14.
[16] J. Liu, Y. Dang, Z. Wang, and T. Zhou, "Relationship between the in-degree and out-degree of WWW," *Physica A*, vol. 371, no. 2, pp. 861–869, 2006.
[17] M. Zanin, P. Cano, O. Celma, and J. M. Buldu, "Preferential attachment, aging and weights in recommendation systems," *International Journal of Bifurcation and Chaos*, vol. 19, no. 02, pp. 755–763, 2009.
[18] A. M. Petersen, W.-S. Jung, J.-S. Yang, and H. E. Stanley, "Quantitative and empirical demonstration of the matthew effect in a study of career longevity," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 18–23, 2011.
[19] C. Watts and N. Gilbert, "Does cumulative advantage affect collective learning in science? an agent-based simulation," *Scientometrics*, vol. 89, no. 1, pp. 437–463, 2011.
[20] R. Kumar, Y. Lifshits, and A. Tomkins, "Evolution of two-sided markets," in *Proceedings of ACM WSDM 2010*, pp. 311–320.
[21] D. Braha, B. Stacey, and Y. Bar-Yam, "Corporate competition: A self-organized network," *Social Networks*, vol. 33, no. 3, pp. 219–230, 2011.
[22] M. Kas, K. M. Carley, and L. R. Carley, "Trends in science networks: understanding structures and statistics of scientific networks," *Social Network Analysis and Mining*, vol. 2, no. 2, pp. 169–187, 2012.
[23] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of ACM SIGKDD 2005*, pp. 177–187.
[24] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the Flickr social network," in *Proceedings of WWW 2009*, pp. 721–730.

APPENDIX

A. Proof of Theorem 1

The basic idea of the proof is that, when the node N_{t+1} enters the network, its probability of linking to the node N_1 is definitely smaller than that to the node N_t. The key observation is that N_1 has at most mt links at the time t (i.e., it attracts all the links of the later nodes). While the linking probability from N_{t+1} to N_t is proportional to $d^\alpha \cdot \Delta t^{-\beta} = m^\alpha$, the linking probability from N_{t+1} to N_1 is asymptotically bounded by $d^\alpha \cdot \Delta t^{-\beta} = (mt)^\alpha \cdot t^{-\beta} = m^\alpha \cdot t^{\alpha-\beta}$. If $\alpha < \beta$, then N_{t+1} is much more likely to link to N_t, instead of N_1 ($m^\alpha \gg m^\alpha \cdot t^{\alpha-\beta}$ when t is large). This implies that N_1 is no longer able to attract new links. N_1 attracts a finite number of links.

B. Proof of Theorem 2

The basic idea of the proof is that, when $\alpha < \beta$, the node N_s needs $\Omega((t-s)^{\beta/\alpha})$ additional links to resist the dominated repulsiveness brought by the age difference. The key observation is that the node N_s can attract more links if we ignore the existences of all the nodes older than N_s. In other words, the upper bound for the degree of the node N_s is the case, where it is regarded as the first node in the growing network. Similar to the proof of Theorem 1, let us focus on the attachment probability for the node that enters the growing network at the time t. While the linking probability from N_t to N_{t-1} is proportional to $d^\alpha \cdot \Delta t^{-\beta} = m^\alpha$, the linking probability from N_t to N_s is at most proportional to $d^\alpha \cdot \Delta t^{-\beta} = d^\alpha \cdot (t-s)^{-\beta}$. To keep the attractiveness of the node N_s, its degree should be larger than $m(t-s)^{\beta/\alpha}$, which can only be brought by its additional initial links. Therefore, at least $\Omega((t-s)^{\beta/\alpha})$ additional links are needed.

C. Proof of Theorem 3

By induction, we now prove that, when $\alpha > \beta + 1.5$, the node N_1 has a degree of at least $c \cdot t$ at the time t. Here, c is a certain constant. This declaration is true, when $t = 1$. Suppose this declaration holds when $t = T$, and then the node N_{T+1} enters the growing network. Note that, the linking probability from N_{T+1} to N_1 is proportional to $d^\alpha \cdot \Delta t^{-\beta} \geq c^\alpha \cdot T^{\alpha-\beta}$. Meanwhile, the linking probability from N_{T+1} to all the other nodes (i.e., N_2, N_3, ..., N_T) is at most proportional to $(2m-c)^\alpha \cdot \int_1^{T-1} \Delta t^{-\beta} \mathrm{d}\Delta t$. This upper bound is obtained by using the average degree of $(2m-c)$ to approximate this linking probability, since older nodes should have larger degrees than younger nodes. When $\beta \geq 0$, we have $\int_1^{T-1} \Delta t^{-\beta} \mathrm{d}\Delta t < T$. Therefore, the condition of $\alpha > \beta + 1.5$ indicates $T^{\alpha-\beta} \gg T$, meaning the node N_1 attracts the most links of the node N_{T+1}. If we set $c \leq m/2$, then the node N_1 has a degree of at least $c \cdot (T+1)$, when $t = T+1$. By induction, the node N_1 has a degree of at least $c \cdot t$ at the time t, when $\alpha > \beta + 1.5$.

D. Proof of Theorem 4

When $\alpha > \beta + 1.5$, Theorem 3 states that the first node N_1 has a degree of $\Theta(s)$, when the node N_s enters the growing network at the time s. Since N_1 has a very large degree, N_s needs some additional initial links to compete with N_1 in the following attachment process. Let us consider the case when the node N_{s+1} enters the network at the time $s+1$. The linking probability from N_{s+1} to N_1 is proportional to $d^\alpha \cdot \Delta t^{-\beta} \in \Theta(s^{\alpha-\beta})$. Therefore, the node N_s needs at least $\Omega(s^{\alpha-\beta})$ additional initial links to break the link monopoly of N_1. Otherwise, the node N_s cannot attract the new links.

E. Proof of Theorem 5

When $|(1-\alpha)C(\alpha,\beta,\xi)| \gg 1$, Eq. 13 can be rewritten as:

$$\begin{aligned} r_g &= \left[\frac{(1+r_i)^{1-\alpha} + (1-\alpha)C(\alpha,\beta,\xi)}{1+(1-\alpha)C(\alpha,\beta,\xi)}\right]^{\frac{1}{1-\alpha}} - 1 \\ &\approx \left[\frac{(1-\alpha)C(\alpha,\beta,\xi)}{(1-\alpha)C(\alpha,\beta,\xi)}\right]^{\frac{1}{1-\alpha}} - 1 = 1 - 1 = 0 \end{aligned} \quad (14)$$

When $|(1-\alpha)C(\alpha,\beta,\xi)| \ll 1$, Eq. 13 can be rewritten as:

$$\begin{aligned} r_g &= \left[\frac{(1+r_i)^{1-\alpha} + (1-\alpha)C(\alpha,\beta,\xi)}{1+(1-\alpha)C(\alpha,\beta,\xi)}\right]^{\frac{1}{1-\alpha}} - 1 \\ &\approx \left[\frac{(1+r_i)^{1-\alpha}}{1+(1-\alpha)C(\alpha,\beta,\xi)}\right]^{\frac{1}{1-\alpha}} - 1 \\ &\approx \left[\frac{(1+r_i)^{1-\alpha}}{e^{(1-\alpha)C(\alpha,\beta,\xi)}}\right]^{\frac{1}{1-\alpha}} - 1 \\ &= (1+r_i)e^{-C(\alpha,\beta,\xi)} - 1 \end{aligned} \quad (15)$$

The above two equations complete the proof of Theorem 5.